QUALITY IMPROVEMENT RESEARCH

Research methods used in developing and applying quality indicators in primary care

S M Campbell, J Braspenning, A Hutchinson, M Marshall

Quality indicators have been developed throughout Europe primarily for use in hospitals, but also increasingly for primary care. Both development and application are important but there has been less research on the application of indicators. Three issues are important when developing or applying indicators: (1) which stakeholder perspective(s) are the indicators intended to reflect; (2) what aspects of health care are being measured; and (3) what evidence is available? The information required to develop quality indicators can be derived using systematic or non-systematic methods. Non-systematic methods such as case studies play an important role but they do not tap in to available evidence. Systematic methods can be based directly on scientific evidence by combining available evidence with expert opinion, or they can be based on clinical guidelines. While it may never be possible to produce an error free measure of quality, measures should adhere, as far as possible, to some fundamental a priori characteristics (acceptability, feasibility, reliability, sensitivity to change, and validity). Adherence to these characteristics will help maximise the effectiveness of quality indicators in quality improvement strategies. It is also necessary to consider what the results of applying indicators tell us about quality of care.

Quality improvement has become a central tenet of health care. It is no longer the preserve of enthusiastic volunteers but part of the daily routine of all those involved in delivering health care, and has become a statutory obligation in many countries. There are numerous reasons why it is important to improve quality of health care, including enhancing the accountability of health practitioners and managers, resource efficiency, identifying and minimising medical errors while maximising the use of effective care and improving outcomes, and aligning care to what users/patients want in addition to what they need.

Quality can be improved without measuring it—for example, by specialist higher educational programmes such as the vocational training scheme for general practice in the UK or guiding care prospectively in the consultation through clinical guidelines. Moreover, there are ways of assessing quality without using hard quantitative measures such as quality indicators—for example, peer review, videoing consultations, patient interviews. Measurement, however, plays an important part in improvement and helps to promote change. Specific measures may, for example, allow good performance to be rewarded in a fair way and facilitate accountability. For this reason much effort has gone into developing and applying measures of quality over the last few decades. The purpose of this paper is to review methods which seek to develop and apply quality indicators.

DEFINING QUALITY INDICATORS

Indicators are explicitly defined and measurable items which act as building blocks in the assessment of care. They are a statement about the structure, process (interpersonal or clinical), or outcomes of care and are used to generate subsequent review criteria and standards which help to operationalise quality indicators (box 1). Indicators are different from guidelines, review criteria, and standards (box 2). Review criteria retrospectively assess care provided on a case-by-case basis to individuals or populations of patients, indicators relate to care or services provided to patients, and standards refer to the outcome of care specified within these indicators. Standards can be 100%—for example, the National Service Framework for coronary heart disease in the UK has set a standard that all patients with diagnosed coronary heart disease should receive low dose (75 mg) aspirin where clinically appropriate. However, care very rarely meets such absolute standards and, in general, standards should be realistic and set according to local context and patient circumstances.

Indicators can measure the frequency with which an event occurred, such as influenza immunisations (activity indicator). However, quality indicators infer a judgement about the quality of care provided. This distinguishes quality indicators from performance indicators, which are statistical devices for monitoring care provided to populations without any necessary inference about quality—for example, they might simply have cost implications. Indicators do not provide definitive answers but indicate potential problems that might need addressing, usually manifested by statistical outliers or perceived unacceptable variation in care. Most indicators have been developed to assess/improve care in hospitals but, increasingly, quality measures are being developed for primary care across Europe.

WHAT SHOULD BE MEASURED?

There are three important issues to consider when developing indicators. Firstly, which stakeholder perspective(s) are the indicators intended to reflect? There are different stakeholders of health...
criterion, and standard

Guideline: systematically developed statements to assist practitioner and patient decisions prospectively for specific clinical circumstances; in essence the "right thing to do." Different perspectives may need different methods of indicator development, particularly as stakeholders have different perspectives about quality of care. Health professionals tend to focus on professional standards, health outcomes, and efficiency. Patients often relate quality to an understanding attitude, communication skills, and clinical performance. Managers’ views are influenced by data on efficiency, patients’ satisfaction, accessibility of care and, increasingly, outcomes. Even if the same aspects of care are assessed, the indicator can be valued differently—for example, health professionals and managers will probably value efficiency differently.

Secondly, which aspects of care should be assessed—processes or outcomes of care? The ultimate goal of the care given to patients can be expressed as outcome indicators which measure mortality, morbidity, health status, health related quality of life, and patient satisfaction. Examples include medical outcomes,5,19 the outcomes utility index,20 the Computerized Needs Orientated Quality Measurement Evaluation System,21 and some of the National Performance Frameworks in the UK.22 Other outcome indicators include user evaluation surveys derived from systematic literature reviews of patient perspectives of health care or outcome indicators developed using focus groups.23 In this way items included in validated patient surveys such as the General Practice Assessment Survey24–26 or Europ27 can be used as quality indicators. One example of such an indicator might be a patient’s capacity to get through to practice staff on the telephone. Structural indicators give information on the practice organisation such as personnel, finances, and availability of appointments.20–23 For example, if a general practice has a car park there should be specified places for disabled parking. There is limited evidence linking structure with outcomes24 although research has suggested, for example, a link between longer consultations and higher quality clinical care.2531 Process indicators describe actual medical care such as diagnoses, treatment, referral, and prescribing.25 Since our focus is on quality improvement, our main interest in this paper is on process indicators because improving process has been described as the primary object of quality assessment/improvement.32

Thirdly, in order to develop indicators researchers need information on structure, process or outcome which can be derived in a number of ways using systematic or non-systematic methods. This information is vital to establish the face or content validity of quality measures (box 3).

RESEARCH METHODS FOR THE DEVELOPMENT OF QUALITY INDICATORS

Non-systematic

Non-systematic approaches to developing quality indicators do not tap in to the evidence base of an aspect of health care; they are based on the availability of data and real life critical incidents. This does not mean that they have no useful role in quality assessment/improvement. Examples include quality improvement projects based on one case study.33 For example, an abortion of a pregnant 13 year old led to a team meeting. Her medical record showed two moments when contraindications could have been discussed. The response was a special clinic hour for teenagers and the development of a quality...
Box 4 What are consensus methods designed to do?

- Enhance decision making, develop policies, and estimate unknown parameters.
- Facilitate the development of quality indicators or review criteria where evidence alone is insufficient.
- Synthesise accumulated expert opinion/professional norms.
- Identify, quantify, and subsequently measure areas where there is uncertainty, controversy, or incomplete evidence.

Systematic: evidence based

Where possible, indicators should be based directly upon scientific evidence such as rigorously conducted (trial based) empirical studies. The better the evidence, the stronger the benefits of applying the indicators in terms of reduced morbidity and mortality or improved quality of care. For example, patients with confirmed coronary artery disease should be prescribed aspirin, unless contraindicated, as there is evidence that aspirin is associated with improved health benefits in patients with coronary heart disease, although the evidence on the exact dose is unclear. McColl and colleagues have developed sets of evidence-based indicators for use by primary care organisations in the UK based on available data.

Systematic: evidence combined with consensus

There are, however, many grey areas of health care for which the scientific evidence base is limited, especially within the generalist and holistic environment of general practice. This necessitates using an extended family of evidence to develop quality indicators, including utilising expert opinion. However, experts often disagree on the interpretation of evidence and rigorous and reproducible methods are needed to assess the level of agreement; in particular, combining expert opinion with available evidence using consensus techniques to assess aspects of care for which evidence alone is insufficient, absent, or methodologically weak. The idea of harvesting professional opinion regarding professional norms of practice to develop quality measures is not new.

Box 4 shows that there are a variety of reasons for developing quality indicators using consensus methods. They also allow a wider proportion of aspects of quality of care to be assessed and thus improved than if indicators were based solely on evidence. Quality indicators abound for preventive care, are patchy for chronic care, and almost absent for acute care in general practice.

Consensus techniques are group facilitation techniques which explore the level of consensus among a group of experts by synthesising and clarifying expert opinion in order to derive a consensus opinion from a group with individual opinions combined into a refined aggregated opinion. Group judgements of professional opinion are preferable to individual practitioner judgements because they are more consistent; individual judgements are more prone to personal bias and lack of reproducibility. Recent examples include quality indicators for common conditions, research on the necessity of process indicators for quality improvement, and a practice visit tool to augment quality improvement.

There are a number of techniques including the Delphi technique, and the RAND appropriateness method which have been discussed elsewhere, and guideline driven indicators using a systematic literature review. The nominal group technique is also used in which a group of experts is asked to generate and prioritise ideas but it is not itself a consensus technique. However, the nominal group technique, supported by postal Delphi, has been used to produce, for example, a national clinical practice guideline in the UK and prescribing indicators.

Delphi technique

The Delphi technique is a structured interactive method involving repetitive administration of anonymous questionnaires, usually across two or three postal rounds. Face to face meetings are not usually a feature. The main stages include: identifying a research problem, developing questionnaire statements to rate, selecting appropriate panelists, conducting anonymous iterative postal questionnaire rounds, feeding back results (statistical, qualitative, or both) between rounds, and summarising and feeding back the findings.

The approach enables a large group to be consulted from a geographically dispersed population. For example, Shield used 11 panels composed of patients, carers, health managers, and health professionals to rate quality indicators of primary mental health care. Optimal size has not been established and research has been published based on samples ranging from 4 to 3000.

The Delphi procedure permits the evaluation of large numbers of scenarios in a short time period. The avoidance of face to face interaction between group members can prevent individuals feeling intimidated and opinions can be expressed away from peer group pressure. However, the process of providing group and, particularly, individual feedback can be very resource intensive. Moreover, absence of any face to face panel discussion opportunity to debate potentially different viewpoints. There is limited evidence of the validity of quality measures derived using the Delphi technique. The Delphi procedure has been used to develop prescribing indicators, managerial indicators, indicators of patient and general practitioner perspectives of chronic illness, indicators for cardiovascular disease, and key attributes of a general practice trainer. The Delphi technique has therefore been used to generate indicators for more than just clinical care.

RAND appropriateness method

This method is a formal group judgement process which systematically and quantitatively combines expert opinion and scientific (systematic literature review) evidence by asking panelists to rate, discuss, and then re-rate indicators. It is the only systematic method of combining expert opinion and evidence. It also incorporates a rating of the feasibility of collecting data, a key characteristic in the application of indicators as discussed below. The main stages include selection of the condition(s) to be assessed, a systematic literature review of the available evidence, generation of preliminary indicators to be rated, selection of expert panels, first round postal survey where panellists are asked to read the accompanying evidence and rate the preliminary indicators, a face to face panel meeting where panellists discuss each indicator in turn, analyses of final ratings, and development of final agreed indicators/criteria. The method has been the subject of a number of critiques.

There have also been used in the UK including the development of review criteria for angina, asthma and diabetes and for 19 common conditions including acute, chronic and preventive care.

The strengths of the RAND method are that panelists meet to discuss and select indicators from the literature, no indicators are discarded between rounds so no potential information is lost and, unlike the standard Delphi technique, panelists are sent a copy of the systematic literature review in addition to the catalogue of indicators. This increases the opportunities for panel members to ground their opinions in the scientific evidence. Research
has also shown that using a higher cut off point for determining consensus within a panel (an overall panel median rating of 8 out of 9) enhances the reproducibility (box 3) of the ratings if a different set of panellists rated the indicators.\(^\text{70}\) Shekelle and colleagues found that, while agreement between panels was weak, in terms of kappa values they had greater reliability than many widely accepted clinical procedures such as reading of mammograms.\(^\text{48}\)

However, the panels inevitably have to be smaller than the Delphi panels for practical reasons, users/patients are rarely involved, the implications of costs are not considered in ratings, and indicators have been limited to clinical care. Moreover, the face to face nature of the discussion can lead to potential intimidation if there are dominant personalities, whereas managers with combined feedback rated lower than managers with manager only feedback.\(^\text{88}\) Different disciplines make systematically different judgements of the acceptability of developed sets in conjunction with a group of lay professionals. The method has evolved within the last decade. Some initial studies assessed the performance of the general practitioner on, for example, threatened miscarriage, asthma and chronic obstructive pulmonary disease where the indicator development was limited to the first round of the procedure.\(^\text{59}\) \(^\text{60}\) Other studies used larger panels to assess key recommendations.\(^\text{56}\) \(^\text{57}\) More recent projects have completed all five rounds—for example, a study in which quality indicators were selected for all 70 guidelines developed by the Dutch College of General Practitioners\(^\text{53}\) or a study on the management of stroke in hospital.\(^\text{59}\)

**Systematic: guideline driven indicators**

Indicators can be based on clinical guidelines.\(^\text{50}\) \(^\text{51}\) \(^\text{52}\) \(^\text{53}\) \(^\text{54}\) \(^\text{55}\) \(^\text{56}\) \(^\text{57}\) Such indicators for general practice have been developed and disseminated widely in the NHS in the UK for four important clinical conditions (diabetes, coronary heart disease, asthma, and depression).\(^\text{58}\) Using methods proposed by AHCPR.\(^\text{59}\) Review criteria were derived from at least one clinical guideline which met a set of quality standards, using structured questions and feedback to test the face and content validity—as well as the feasibility—of the criteria with a panel of over 60 general practitioners.

Hadorn and colleagues\(^\text{60}\) described how 34 recommendations in a guideline on heart failure were translated into eight review criteria. Because review criteria must be specific enough to assure the reliability and validity of retrospective review, they used two selection criteria to guide whether each recommendation was specific or general. They demonstrated some important aspects of criteria development from guidelines, in particular the need to be very detailed and specific in the criterion, even though the guideline recommendation is less specific and deemed adequate.

Review criteria derived directly from a clinical practice guideline are now part of NHS policy in England and Wales through the work of the National Institute of Clinical Excellence (NICE). Each published summary clinical guideline is accompanied by a set of review criteria which are intended to be used by clinical teams, and the results are externally assessed by the Commission for Health Improvement—for example, in relation to type 2 diabetes.\(^\text{61}\)

*Table 1* Guideline driven indicators developed using an iterated consensus rating procedure

<table>
<thead>
<tr>
<th>Round</th>
<th>Aim</th>
<th>Undertaken by</th>
<th>Criteria used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1: Pre-selection</td>
<td>Selecting key recommendations</td>
<td>Small group of quality indicators developers (3 persons)</td>
<td>Outcome of care: Patients’ health (morbidity, mortality, health status) Cost Patients’ health Cost Sensitivity to change Availability of data Kappa, rho</td>
</tr>
<tr>
<td>Round 2: Rating and adding key recommendations</td>
<td>Rating and adding key recommendations</td>
<td>Expert panel (8–10 persons)</td>
<td>Cut off score: mean above mid of rating scale Agreement among 80% of the panel members Face validity</td>
</tr>
<tr>
<td>Round 3: Reliability</td>
<td>Determining inter- and intra-rater reliability</td>
<td>Expert panel for the rating</td>
<td>Research team for the analyses</td>
</tr>
<tr>
<td>Round 4: Potential indicators</td>
<td>Getting set of potential indicators</td>
<td>Research team</td>
<td>Face validity</td>
</tr>
<tr>
<td>Round 5: Reflection</td>
<td>Acceptability of indicators</td>
<td>Research team</td>
<td>Laymen professionals</td>
</tr>
</tbody>
</table>

**Factors influencing the development of quality indicators using a consensus technique**

Many factors influence ratings in a consensus method, especially group composition as groups composed of different stakeholders rating the same statements produce different ratings.\(^\text{26}\) \(^\text{59}\) \(^\text{79}\) \(^\text{80}\) \(^\text{81}\) For example, group members who use, or are familiar with, the procedures being rated are more likely to rate them higher.\(^\text{59}\) \(^\text{79}\) \(^\text{80}\) \(^\text{81}\) Moreover, panel members from different disciplines make systematically different judgements and feedback from mixed disciplines may influence ratings. For example, a Delphi composed equally of health physicians and managers found that the physicians who had overall feedback, including that of the managers, rated indicators higher than the physicians who had physician only feedback, whereas managers with combined feedback rated lower than managers with manager only feedback.\(^\text{84}\)

Ongoing work has provided qualitative evidence of factors which influence individual panellists’ ratings in a consensus technique rating aspects of the quality of primary mental health care in a two round postal Delphi.\(^\text{53}\) This research used in depth qualitative interviews with panellists from patient, managerial, and professional panels to identify factors which had influenced panellists’ ratings. It concluded that many factors influenced the ratings of the different stakeholder groups (box 5).

**Research methods on the application of indicators**

Measures derived using expert panels and guidelines have high face validity and those based on rigorous evidence possess high content validity. However, this should be a minimum prerequisite for any quality measure and subsequent developmental work is required to provide empirical evidence, as far as possible, of acceptability, feasibility, reliability, sensitivity to change, and predictive validity (box 3).\(^\text{68}\) \(^\text{91}\) \(^\text{82}\)
Acceptability
The acceptability of the data collected using a measure will depend upon the extent to which the findings are acceptable to both those being assessed and those undertaking the assessment. For example, the iterated consensus rating procedure consults lay professionals as to the acceptability of indicators (table 1). Campbell and colleagues conducted a quality assessment in 60 general practices in England but only used quality indicators rated acceptable and valid by the nurses and doctors working in the practices.\(^9\)

Feasibility
Information about the quality of services is often driven by data availability rather than by epidemiological and clinical considerations.\(^9\) Quality measurement cannot be achieved without accurate and consistent information systems.\(^9\) Current administrative data, both at the macro (health authority or “large organisation”) and micro (individual medical records) levels, are constrained by inconsistent and often unreliable data.\(^9\) Medical records are a poor vehicle for collecting data on preventive care and the recording of symptoms.\(^9\)

In addition, aspects of care being assessed by quality indicators must relate to enough patients to make comparing data feasible. For example, a clinical audit of angina care excluded 10 criteria rated necessary by an expert panel to provide quality of care because they related to less than 1% of a sample of over 1000 patients in 60 general practices in England.\(^9\)

Reliability
Indicators should be used to compare organisations/practitioners with similar organisations/practitioners, or confounding factors such as socioeconomic and demographic factors, as well as factors outside the control of practitioners, should be taken into account (that is, compare like with like or risk/case mix adjust). This is because the environment in which an organisation operates affects the care provided. Examples include admission rates or surgery rates. Indicators must also have explicit exclusion and inclusion criteria for applying the indicator to patients—for example, age ranges, co-morbidities, case mix, and clinical diagnoses.

The inter-rater reliability of an indicator can also be tested when applying indicators. For example, in a study of over 1000 patients with diabetes two raters abstracted data separately (but on the same day) for 7.5% of all patient records and found that five criteria out of 31 developed using an expert panel were excluded from analyses due to poor agreement.\(^9\)

Sensitivity to change
Quality measures must be capable of detecting changes in quality of care in order to discriminate between and within subjects.\(^9\) This is an important and often forgotten dimension of Lawrence’s definition of a quality indicator.\(^9\)

Validity
There has been little methodological scrutiny of the validity of consensus methods.\(^9\) The Delphi technique\(^9\) and the RAND method\(^9\) have both been criticised for a lack of evidence of validity. While the issue has received more attention in recent years,\(^9\) there is little evidence for the validity of the Delphi method in developing quality indicators.

Content validity of indicators generated using consensus techniques
Content validity in this context refers to whether any indicators were rated by panels contrary to known results from randomised controlled trials. There is evidence for the content validity of indicators derived using the RAND method.\(^9\)

Predictive validity
There is evidence of the predictive validity of indicators developed using the RAND method.\(^9\) For example, Kravitz and colleagues studied a cohort of persons who had undergone coronary angiography. Patients were retrospectively classified as to whether coronary revascularisation was “necessary” or “not necessary” according to the review criteria, and outcomes at year 1 were measured. Patients meeting the “necessary” criteria for coronary revascularisation who did not receive it were twice as likely to have died at 1 year as those who did receive “necessary” revascularisation. Hemingway et al\(^9\) found substantial underuse of coronary revascularisation among UK patients who were considered appropriate for these procedures and underuse was associated with adverse clinical outcomes on the basis of the ratings of an expert panel.

**Box 5 Factors influencing indicators rated valid in a Delphi technique**\(^9,93\)

- Composition of the panel
- Inclusion of patient derived (focus groups) indicators
- Inclusion of indicators based on "grey" literature
- Inclusion of multiple stakeholders (e.g. patients, carers, managers, health professionals)
- Characteristics of individual panelists (e.g. political perspective, familiarity with research)
- Rating process (e.g. 9 point scale, feedback used)
- Panellists’ experience and expectations of the care provision being rated
- Panellists’ perspective of the model of care provision
- Panellists’ perspective of their locus of control to influence care

**USING DATA GENERATED BY APPLYING QUALITY INDICATORS**
Data generated using quality indicators can be used for a variety of purposes—for example, to monitor, reward, penalise, or compare care provision (perhaps using league tables or public release of data) or as part of a quality improvement strategy. Simply measuring something will not automatically improve it. Indicators must be used within coherent systems based approaches to quality improvement.\(^9\) The interpretation and usage of such data is more of a political or resource issue than a methodological or conceptual one.

The provenance of the indicators is important when applying them. Indicators derived from informal consensus procedures with little evidence underlying them might be useful as educational guidelines. However, the best indicators for public disclosure, for use in league tables, or for attaching financial incentives are those based solely on scientific evidence, for which the implications of applying the indicator and any relative judgements that are be inferred about the results can be confidently predicted. Indicators derived from consensus methods which systematically combine evidence and opinion may also be disclosed, but perhaps with more provisos. Indicators developed by well respected experts using a systematic method might also have high credibility when used for professional development.

**CONCLUSION**
It may never be possible to produce an error free measure of quality, but measures should adhere, as far as possible, to some fundamental a priori characteristics in their development (face/content validity) and application (acceptability, feasibility, reliability, sensitivity to change, predictive validity). Adherence to these characteristics will help maximise the effectiveness of quality indicators in quality improvement strategies. This is most likely to be achieved when they are derived from rigorous scientific evidence. However, evidence in health care is often absent. We believe that using consensus techniques—which systematically combine evidence and opinion—and guideline driven approaches facilitates quality...
improvement. They allow a significantly broader range of aspects of care to be assessed and improved than would be the case if quality indicators were restricted to scientific evidence. It is important that such methods of development continuously improve and seek to incorporate advances in the evidence base of health care. However, it may be that research has reached a peak in developing indicators. There is much less research on the application of indicators and their reliability, validity, and effectiveness in quality improvement strategies, how indicators can be used to improve care, and how professionals/service users can be helped to be more engaged with the development and use of indicators. Introducing strategies for quality improvement based on quality indicators does not make them effective and successful.

Key messages

- Most quality indicators have been developed in hospitals but they are increasingly being developed for primary care in Europe and the USA.
- Most research has focused on the development rather than the application of indicators.
- Quality indicators should be based on rigorous scientific evidence if possible. However, evidence in health care is often absent, necessitating the use of other methods of development and the use of consensus techniques, such as the Delphi technique and the RAND appropriateness method, which combine expert opinion and available evidence and indicators based on clinical guidelines.
- While it may never be possible to produce an error free measure of quality, measures should adhere, as far as possible, to some fundamental a priori characteristics, namely, acceptability, feasibility, reliability, sensitivity to change, and validity.
- The way in which indicators are applied is as important as the method of development.

Authors' affiliations

S M Campbell, M Marshall, National Primary Care Research and Development Centre, University of Manchester, Manchester M13 9PL, UK
A Hutchinson, University of Sheffield, Section of Public Health, Sheffield S1 4DA, UK
J Bransenpenning, UMC St Radboud, WOK, Centre for Quality of Care Research, Postbus 9101, 6500 HB Nijmegen, The Netherlands

REFERENCES


www.qualityhealthcare.com


Research methods used in developing and applying quality indicators in primary care

S M Campbell, J Braspennings, A Hutchinson and M Marshall

Qual Saf Health Care 2002 11: 358-364
doi: 10.1136/qhc.11.4.358

These include:

References
This article cites 81 articles, 35 of which you can access for free at:
http://qualitysafety.bmj.com/content/11/4/358#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/